JOIV : Int. J. Inform. Visualization, 7(1) - March 2023 131-138

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Hybrid Approach with Distance Feature for Multi-Class Imbalanced Datasets

Hartono [a,*], Erianto Ongko [b]

[a] Department of Computer Science, Universitas Potensi Utama, Medan, 20241, Indonesia
[b] Department of Informatics, Akademi Teknologi Industri Immanuel, 20114, Medan, Indonesia
Corresponding author: *hartonoibbi@gmail.com

***Abstract*—The multi-class imbalance problem has a higher level of complexity when compared to the binary class problem. The difficulty is due to the large number of classes that will present challenges related to overlapping between classes. Many approaches have been proposed to deal with these multi-class problems. One is a hybrid approach combining a data-level approach and an algorithm-level approach. This approach is done by the ensemble on the classifier and also oversampling on the minority class. SMOTE is an oversampling method that provides good performance, but this method is necessary to determine the best sample used in the interpolation process to generate new samples. The need for determining the best sample is related to the overlap between classes that always accompanies the multi-class imbalance problem. The existence of overlap requires efforts to determine the safe region to synthesize the sample in the oversampling process in SMOTE. The safe region is considered the best for synthesizing samples due to the lower tendency of overlapping. It can be done by constructing distance features to determine the safe region. The sample with the best distance and the lowest imbalance ratio will be selected as a sample in the over-sampling process with SMOTE. The main contribution of this research is the proposed method of Hybrid Approach with Distance Feature so that it can determine safe samples, with the main advantage being in addition to handling multi-class imbalances, it is also better for handling overlapping. The results of this study will be compared with Multiple Random Balance (MultiRandBal) which performs a random oversampling process. The results showed that the Augmented R-Value, Class Average Accuracy, Class Balance Accuracy, and Hamming Loss obtained in this method was better than the random oversampling process. These results also show that the Hybrid Approach with Distance Feature provides better results in handling multi-class imbalances when compared to MultiRandBal.**

***Keywords*— Multi-class imbalance; overlapping; hybrid approach; distance feature; SMOTE.**

Manuscript received 17 Oct. 2022; revised 14 Jan. 2023; accepted 25 Jan. 2023. Date of publication 31 Mar. 2023.

I. INTRODUCTION

In recent years, the issue of class imbalance has become a concern for nearly all machine learning algorithms [1]. Datasets with imbalance problems tend to present greater challenges in the discussion of machine learning, especially when we talk about the quality and accuracy of classification [2]. In the real world, datasets frequently suffer from class imbalance issues, as evidenced by the presence of classes with significantly more instances than other classes [3]. The classification accuracy results in the majority class will be better when compared to the minority class [4]. In datasets that experience class imbalance, there is a tendency that the minority class is a class with information that tends to be more interesting than the majority class [5]. Class imbalance handling can increase overall classification accuracy [6].

Researchers have realized that the problem of class imbalance is a difficult problem to handle, so several methods have been proposed to deal with this problem. A number of these techniques can be grouped into 3 (three) groups: data-level, algorithm-level, and hybrid approach [7]. The data-level approach is put into practice by either undersampling the majority class or oversampling the minority class. The algorithm level is implemented as bagging and boosting [8]. In comparison, the hybrid approach implements a combination of data-level and algorithm-level[9]. The combination of data-level and algorithm-level approaches found in the hybrid approach can provide better results in handling class imbalance [10].

The oversampling method has been used by many researchers, with the largest number of oversampling methods used being the Synthetic Minority Oversampling Technique (SMOTE) [11]. It should be noted that the effort required to

deal with multi-class imbalance is much greater than that of binary class [12]. Many minority-one majority, one minority-many majority, and many minority-many majority are examples of multi-class problems [13]. There are several types of examples in multi-classes: Safe Examples, Rare Examples, Borderline Examples, and Outliers[14].

SMOTE will perform the interpolation process to generate new instances based on existing samples [15]. The interpolation process can be done by selecting a sample at random or by collecting information from the sample. Information from samples that need to be collected is related to the distance of the sample to the sparse area and the decision boundary [16]. It should be noted that there are several minority instances selected from the instances that are in the safe area based on the determination with the nearest neighbor can give better results on the data generation process [2]. Inaccuracy in selecting samples for the interpolation process in SMOTE can also result in overlapping [17].

Class overlapping is already at risk even in balanced conditions and becomes riskier in conditions accompanied by class imbalance and increasingly affects accuracy in multi-class imbalance conditions [18]. Overlapping becomes difficult to handle and dangerous in multi-class because of the increasingly blurred boundaries between existing classes, resulting in inaccurate classification results[19].

Safe examples in multi-class dataset are the most easily identified and classified samples [20]. Borderline examples are samples that lie within the boundary between several classes so that there may be overlapping samples between several classes[21]. There are also samples located in areas where many samples prevail from other classes (outliers) [22]. Finally, there are several samples located far from their group and form groups in other classes called rare samples [23].

Problems that often occur in borderline, rare, and outliers are related to overlapping[24]. The safe region is considered the best for synthesizing samples due to the lower overlapping tendency. To ensure the determination of samples that are in the safe region, it can be done by constructing distance features[25]. For samples whose imbalance ratio is not greater than the mean imbalance ratio, it can be directly selected as a sample in SMOTE, while the imbalance ratio is greater than the mean imbalance ratio, so it is necessary to determine the distance. The sample with the best distance and the lowest imbalance ratio will be selected as a sample in the over-sampling process with SMOTE[26]. The results of the research conducted by[27] support that Distance Features are worthy of consideration for determining the learned boundary because if a sample has a simultaneously learned distance far from the nearest neighbor, misclassification may occur.

A variety of approaches have been proposed to address the multi-class imbalance and overlapping issues. Among them is the Multi-Class Combined Cleaning and Re-sampling (MC-CCR) method, and MC-CCR requires good cleaning and parameter settings. In addition, the MC-CCR is also difficult if the selected sample is a noise sample[28].

The solution to overcome outliers and rare samples is to balance the training dataset as done by[29] using Interquartile Range (IQR) algorithm. However, the disadvantage is that it relies on the extreme value of outliers, so if the values in outliers and rare are not too much different from instances in other groups, this method cannot function properly. The main focus should be on how to generate samples that come from safe samples[30].

The method that provides good accuracy in handling class imbalance is Multiple Random Balance (MultiRandBal)[31], but with random sample determination, it is likely to be stuck with overlapping conditions. The overlapping conditions need to be well understood because there is a tendency to cause high accuracy in one class, which can reduce accuracy in other classes.

Based on what was conveyed by a number of researchers, it appears that the major concern of SMOTE is how to determine the sample to be used in the interpolation process to generate new samples. The need for determining the best sample is related to the overlap between classes that always accompanies multi-class imbalance. Class imbalance affects the classification results, but the problem of overlapping cannot be ruled out because it greatly affects the accuracy of the classification results [32]. In addition, the determination of samples in SMOTE that disregards the Safe Sample tends to result in the omission of crucial information regarding positive samples, since the number of occurrences is extremely small [33]. Moreover, noise samples are the primary cause of misclassification in datasets [34].

This paper's main contribution is to the proposed method using Hybrid Approach with Distance Feature. The main advantage is to increase SMOTE's ability to determine safe samples. Determination of safe samples is very important in handling multi-class imbalances because often, multi-class imbalances are accompanied by overlapping. By using the Safe Sample, the benefits obtained, in addition to better handling of multi-class imbalance, are also handling of overlapping, so that it is hoped that the accuracy and classification results obtained are also better. The handling of Multi-class imbalance and overlapping, which is carried out by using Safe Sample selection with Distance Feature in the Hybrid Approach is proposed in this study and is the novelty of this research.

The implementation of this research will be carried out to measure the overlapping using Augmented R-Value for Multi-Class, the average accuracy of the class using the Class Average Accuracy parameter, the balance of accuracy using the Class Balance Accuracy, and also the misclassification of the group label using Hamming Loss.

The results of the Hybrid Approach with Distance Feature will be compared to the results of MultiRandBal in this study.

II. Materials and Method

A. Hybrid Approach

The following pseudocode demonstrates how the Hybrid Approach works [35].

$Input: D_T = \{x_1, x_2, \dots, x_n\} // Training\ Dataset$
$N = Number\ of\ Classifier$
$Output: Classification\ Prediction\ P$
$Method:$
$Step\ 1\ Preprocessing\ using\ Preprocessing\ Method$
$Step\ 2\ For\ i = 1\ to\ N\ do$
$i. Apply\ Machine\ Learning\ Classification\ Algorithm\ on\ The\ Attributes\ of\ D_T$
$ii. Obtain\ Classification\ Prediction\ P_i\ from\ machine\ learning\ classification\ algorithm$

End For
Step 3 For $i = 1$ *to* n
Apply processing using bagging, boosting or sampling
End For

Through the pseudocode shown, it is clear that the Hybrid Approach will determine the number of classifiers. The process generally consists of 2 stages: applying machine learning for classification. Then if there is a class imbalance problem, it will be combined with some bagging, boosting, and sampling to solve the class imbalance problem.

B. Synthetic Minority Over-sampling Technique (SMOTE)

The following pseudocode demonstrates how the SMOTE works [36].

Input: $X_{minor}, N_{percent}, K$
Function SMOTE $(X_{minor}, N_{percent}, K)$
$X_{SMOTE} \leftarrow \{\ \}$
for $i \leftarrow 1$ *to* $len(X_{minor})$ *do*
  $nn \leftarrow K$ *Nearest Neighbors* $(X_i, N_{Percent}, K)$
  $p \leftarrow [Npercent/100]$
  while $p! = do$
    $X_{neighbour} \leftarrow select\ random\ (nn)$
    $X_{SMOTE} \leftarrow X_i + rand(0,1) * |X_{neighbour} - X_i|$
    $p \leftarrow p - 1$
  end while
end for
return $\leftarrow X_{SMOTE}$

The selection of minority samples marks the beginning of the SMOTE process, as can be seen from the pseudocode. Then the process will continue with determining the classifier used to carry out the oversampling process. Through the efforts made in the oversampling process, it is expected that the number of samples in the minority class can increase.

C. IRlabel and MeanIR

The balance ratio for multiple classes is expressed by an IRlabel, which can be determined using Equation 1[37].

$$IRlabel_i = \frac{Y_{max}}{\sum_{j=1}^{s} y_{ji}}, where\ Y_{max} = \max_{1 \le j \le b} \left\{ \sum_{i=1}^{s} y_{ji} \right\} \quad (1)$$

where $Y \in \{0,1\}^{sxb}$, *is the number of instance*

While the average value of the imbalance ratio can be determined using Equation 2[37].

$$meanIR = \frac{1}{b} \sum_{i=1}^{b} IRlabel_i \quad (2)$$

The greater the value of IRlabel and meanIR, the greater the imbalance.

D. Distance Feature

Mishra and Singh [25] have stated the importance of determining the distance in determining the safe region. The pseudocode for determining the distance feature is as follows.

Input: $T = (X, Y)$ *be the input dataset, where* $X \in R^{sxf}, Y \in \{0,1\}^{sxb}$,
$F = \{F_1, F_2, \dots, F_f\}$ *is the feature set and* $L = \{L_1, L_2, \dots, F_b\}$ *is the set of class labels.* X_u *is the unseen instance, whose label need be predicted*
Output: The predicted label vector Y_u

$Y_u \leftarrow \emptyset$
for $k \leftarrow 1$ *to* b *do*
  if $IRlabel_k > meanIR$ *then*
    $x_k^+ \leftarrow \{X_i : X_i \in X\ and\ y_{ik} = 1\}$
    $d_k \leftarrow Distances\ of\ X_u\ from\ all\ the\ instances\ in\ X_k^+$
    $y_k \leftarrow h_k(d_k)//$
Predict k^{th} *label with transformed feature space*
  else
    $y_k \leftarrow h_k(x_u)//$
Predict k^{th} *label with original feature space*
  end if
  $y_u \leftarrow y_u \cup y_k$
end for
return y_u

Based on the pseudocode, it can be seen that the process of determining the predicted label, which is declared as y_u with is based on several conditions where for the sample with $IRlabel_k < meanIR$, the sample will automatically become the predicted sample. As for the sample with $IRlabel_k > meanIR$ that will be another step for determining the distance. The sample with the best distance value will be selected as the predicted sample. Determining the distance can be done using Euclidean Distance, as seen in Equation 3.

$$d_k = \sqrt{\sum_{i=1}^{k} (X_u - X_k^+)^2} \quad (3)$$

E. Augmented R-Value for Multi-Class

The overlapping level is expressed in Augmented R-Value. In multi-class problems Augmented R-Value is measured for each class. The higher the Augmented R-Value value, the higher the overlapping level. The Augmented R-Value can be measured by Equation 4.

$$R_{aug}(D[V]) = \frac{\sum_{i=0}^{k-1} |C_{k-1-i}| R(C_i)}{\sum_{i=0}^{k-1} |C_i|} \quad (4)$$

Where $C_0, C_1, \dots, C_{k-1}$ are k class labels with $|C_0| \ge |C_1| \ge \cdots \ge |C_{k-1}|$ and $D[V]$: Dataset D restraining predictors in set *V*.

F. Classifier Performance

Classifier Performance will be measured using the Class Average Accuracy, Class Balance Accuracy, and Hamming Loss. The confusion matrix presented in Table I is intended to measure classifier performance[38].

TABLE I
CONFUSION MATRIX

		Predicted (Classified)	
		Positive	**Negative**
Real	Positive Samples	TP	FN
	Negative Samples	FP	TN

Let $TP_k, FP_k, TN_k, and\ FN_k$ for multi-class imbalance can be calculated as follows[25].

$$TP_k = \left| \{ y_{kj} : y_{kj} = 1\ and\ y'_{kj} = 1 \} \right| \quad (5)$$

$$FP_k = \left| \{ y_{kj} : y_{kj} = 1\ and\ y'_{kj} = 0 \} \right| \quad (6)$$

$$TN_k = \left| \{ y_{kj} : y_{kj} = 0\ and\ y'_{kj} = 0 \} \right| \quad (7)$$

$$FN_k = \left| \{ y_{kj} : y_{kj} = 0\ and\ y'_{kj} = 1 \} \right| \quad (8)$$

Class Average Accuracy, Class Balance Accuracy, and Hamming Loss can be calculated using the following equation [39], [40], [41].

$$AvAcc = \frac{1}{C}\sum_{k=1}^{C}\frac{TP_k+TN_k}{TP_k+TN_k+FP_k+FN_k} \quad (9)$$

$$CBA = \frac{\sum_{i}^{k}\frac{c_{ii}}{\max(c_i,c_i)}}{k} \quad (10)$$

$$Hamming\ Loss\ (h) = \frac{1}{r}\sum_{k=1}^{r}\frac{FP_k+FN_k}{TP_k+FP_k+TN_k+FN_k} \quad (11)$$

The average accuracy obtained by dividing the sum of True Positive and True Negative by the sum of True Positive, True Negative, False Positive, and False Negative is referred to as the class average accuracy. where CBA is defined as an overall accuracy measure derived from the sum of individual class assessments. Hamming Loss declared a label group that was not properly classified. The lower the Hamming loss value, the better the performance classifier.

G. Proposed Method / Algorithm

Figure 1 depicts the research phase.

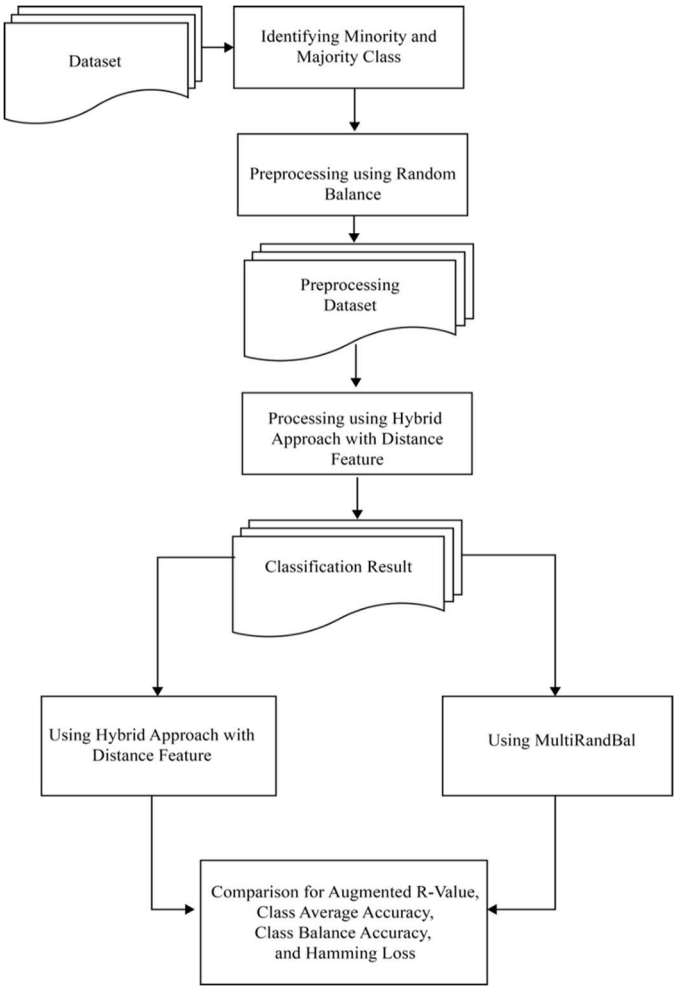

Fig. 1 Research Stage

Figure 1 shows that the research stages can be divided into 2 (two), namely: the preprocessing stage and the processing stage. The process begins by determining the majority and minority classes. The preprocessing stage is carried out using the Random Balance Ensemble Method. The result of the preprocessing stage is in the form of preprocessing the dataset, which will then undergo the processing stage. This processing stage is the main stage in handling class imbalance. This processing stage is carried out using a combination of data-level and algorithm-level (hybrid approach). To improve the performance of the hybrid approach, the selected sample is a sample from a safe region. The determination of the safe region is based on the distance feature. The final results will be compared with the Hybrid Approach method without distance features. Penggunaan Distance feature dimaksudkan untuk menentukan safe sample. The distance determination is done with conditions where the sample with $IRlabel_k < meanIR$, the sample will automatically become the predicted sample. As for the sample with $IRlabel_k > meanIR$ that will be another step for determining the distance. The sample with the best distance value will be selected as the predicted sample.

Based on Figure 1, it can be seen that the contribution of this study is the determination of samples originating from the Safe Region by using the distance feature determination in the Hybrid Approach, with the main advantage being better classification accuracy and also handling overlapping.

1) Preprocessing Using Random Balance Ensemble Method: The preprocessing stage, which is the earliest stage in handling class imbalance, can be seen in how it works in general as in the following pseudocode.

Input: $Set\ S\ of\ Samples. Total\ Size\ totalSize, Number\ of\ Majority\ S_N, Number\ of\ Minority\ S_P$
Output: $New\ set\ S'\ of\ Samples$
1: $totalSize \leftarrow |S|$
2: $S_N = \{(x_i, y_i) \in S | y_i = -1\}$
3: $S_P = \{(x_i, y_i) \in S | y_i = +1\}$
4: $majoritySize \leftarrow |S_N|$
5: $minoritySize \leftarrow |S_P|$
6: $Calculate\ the\ size\ of\ Majority\ Class\ from\ S_N$
7: $Calculate\ the\ size\ of\ Minority\ Class\ from\ S_P$
8: $newMajoritySize \leftarrow$
$Random\ Integer\ Between\ 2\ and\ totalSize - 2$
9: $newMinoritySize \leftarrow totalSize - newMajoritySize$
10: $if\ newMajoritySize < majoritySize\ then$
11: $S' \leftarrow S_P$
12: $S'\ will\ fill\ with\ a\ random\ instance\ from\ S_N$
13: $Create\ newMinoritySize - minoritySize\ artificial$
14: $else$
15: $S' \leftarrow S_P$
16: $S'\ will\ fill\ with\ a\ random\ instance\ from\ S_P$
17: $Create\ newMajoritySize - majoritySize\ artificial$
18: $end\ if$
19: $return\ S'$

The preprocessing stage starts with figuring out how big the majority class and minority class are, as can be seen from the pseudocode. Then the process will continue with the synthesis of newMajorityClass and newMinority Class. If the synthesis results show that newMajorityCass has a smaller number of samples than MajorityClass, a random number of samples will be removed from newMinorityClass and otherwise, newMajorityClass will be discarded randomly.

2) Processing Using SMOTE and Distance Features: The pseudocode of the processing stage is as follows.

Input: $Set\ S\ of\ Samples$
Output: $New\ set\ S'\ of\ Samples$

1: $totalSize \leftarrow |S|$
2: $Determine\ k\ as\ the\ number\ of\ Nearest\ Neighbor$
3: $For\ All\ Samples\ in\ S\ do$
4: $Determine\ the\ borderline\ of\ Positive\ or\ Minority\ Class\ as\ E_O C_t^+$
5: $Determine\ the\ borderline\ of\ Negative\ or\ Majority\ Class\ as\ E_O C_t^-$
6: $End\ For$
7: $For\ k \leftarrow 1\ to\ number\ os\ samples\ in\ E_O C_t^+\ do$
8: $Calcuate\ the\ cn(e)_k\ as\ neighborhood\ value\ for\ each\ sample$
9: $Order\ Ascending\ the\ Sample\ According\ to\ the\ cn(e)_k$
10: $End\ For$
11: $Building\ a\ candidate\ ensemble\ for\ Safe, Borderline, Rare, and\ Outlier\ according\ to\ k\ value$
12: $Take\ a\ candidate\ ensemble\ for\ Safe, Borderline, Rare, and\ Outlier\ according\ to\ k\ value$
13: $For\ k\ \leftarrow 1\ to\ All\ Samples\ in\ E_O C_t^- do$
14: $Take\ a\ candidate\ ensemble\ to\ X_k^-$
15: $End\ For$
16: $For\ k \leftarrow 1\ to\ All\ Samples\ in\ E_O C_t^+ do$
17: $Take\ a\ candidate\ ensemble\ to\ X_k^+$
18: $End\ For$
19: $For\ k \leftarrow 1\ to\ number\ of\ Samples\ in\ X_k^+\ do$
20: $Calculate\ IRlabel_k\ of\ samples\ k\ by\ Eq.1$
21: $Calcuate\ MeanIR\ by\ Eq.2$
22: $If\ IRlabel_k > MeanIR$
23: $X_k' \leftarrow Calculate\ Distance\ of\ X'$ using Eq.2
24: $X_k^{smote} \leftarrow smote\ (X_k', P)$
25: $S_P' \leftarrow X_k' \cup X_k^{smote}$
26: $else$
27: $S_p' \leftarrow X_k^+$
28: $End\ if$
29: $S' \leftarrow X_k^- \cup S_P'$
30: $End\ For$

The processing steps are carried out using a Hybrid Approach with a Distance Feature. The process begins with determining the ensemble candidates for Safe, Borderline, Rare, and Outlier. Then determine all candidate samples that come from positive samples (minority class) and negative samples (majority class). For each sample that comes from positive samples, determine the IRLabel and MeanIR. If the IRlabel value is greater than MeanIR, then the distance is determined for each sample where the sample with the best distance will be selected and if the IRlabel value is smaller than MeanIR, the existing sample will be directly selected.

III. Results and Discussion

A. Dataset Description

The dataset used in this study was obtained from the UCI Machine Learning Repository [42], as seen in Table II.

TABLE II
DATASET DESCRIPTION

Dataset	No. of Examples	No. Of Attributes	Class Distribution	No. of Class	IR
Balance	625	4	49;288;288	3	5.9
Contraceptive	1473	9	629;333; 511	3	1.9
Dermatology	366	34	112;61;72; 49;52;20	6	5.6
Ecoli	327	6	143;77;35; 20;52	5	7.15
Glass	214	9	70;76;17; 13;9;29	6	8.4
Page-Block	5473	10	4913;329; 28;88;115	5	175.4

In Table II, the dataset used has various number of instances, a number of attributes, and also with various imbalance ratios. With variations in the dataset, it is hoped that the results obtained can describe the performance of the proposed method.

B. Experimental Setup

The testing process is carried out on the selected dataset. The process is carried out to test the level of overlapping that occurs because it is understood that by selecting samples in the safe region, the overlapping that occurs should be reduced. Therefore, this experiment was conducted to measure the Augmented R-Value for Multi-Class. The results of this experiment will be validated using a stratified k-fold (k=10).

The performance testing on the classifier is based on the value of class average accuracy, class balance accuracy, and Hamming Loss. These three parameters are intended to provide an overview of the level of accuracy and classification errors that occur. These three parameters can provide a complete picture of the performance of the classifier. Just like in overlapping, the validation of the test results will also use a stratified k-fold (k=10).

C. Testing Result

To obtain Augmented R-Value for Multi-Class and Class Average Accuracy, the first test was carried out. The test results can be seen in Table III.

TABLE III
TESTING FOR AUGMENTED R-VALUE FOR MULTI-CLASS AND CLASS AVERAGE ACCURACY

Dataset	Hybrid Approach with Distance Feature		MultiRandBal	
	Augmented R-Value for Multi-Class	Class Average Accuracy	Augmented R-Value for Multi-Class	Class Average Accuracy
Balance	0.243	0.973	0.265	0.967
Contraceptive	0.251	0.941	0.252	0.937
Dermatology	0.275	0.873	0.301	0.865
Ecoli	0.261	0.965	0.273	0.956
Glass	0.268	0.937	0.314	0.921
Page-Block	0.271	0.861	0.341	0.792

For the handling of overlapping, it can be stated that based on what is shown in the research results using Augmented R-Value for Multi-Class value, the results obtained are Hybrid Approach with Distance Feature which obtains better results when compared to MultiRandBal. This shows that the selection of samples in the Safe Region, carried out through distance feature measurements, obtains good results so that the level of overlap tends to be lower. Through the results of the study, it is also known that the things that most influence overlapping besides the imbalance ratio are the number of classes and the number of attributes.

The test results that have previously been shown make it increasingly clear that the results obtained are related to the Class Average Accuracy is strongly influenced by the Balance Ratio and the number of attributes. This applies to both the Hybrid Approach with Distance Feature and MultiRandBal. This condition is indicated by the lower Class Average Accuracy value in the Dermatology and Page-Block datasets compared to other datasets. In this test, the Hybrid Approach with Distance Feature gives better results than MultiRandBal.

Based on the test results, it can also be seen that there is a tendency that more Augmented R-Value for Multi-Class values can be obtained, which indicates that the higher the overlapping level, the lower the accuracy obtained. This shows that the effect of overlapping is quite large on the accuracy of the classification results, so it needs serious attention.

The second test was conducted to obtain Class Balance Accuracy and Hamming Loss. The test results can be seen in Table IV.

TABLE IV
TESTING FOR CLASS BALANCE ACCURACY AND HAMMING LOSS

Dataset	Hybrid Approach with Distance Feature		MultiRandBal	
	Class Balance Accuracy	Hamming Loss	Class Balance Accuracy	Hamming Loss
Balance	0.973	0.087	0.961	0.121
Contraceptive	0.861	0.106	0.797	0.117
Dermatology	0.837	0.217	0.807	0.256
Ecoli	0.965	0.103	0.966	0.118
Glass	0.897	0.127	0.861	0.158
Page-Block	0.832	0.176	0.732	0.271

Based on the test results, it can be seen that the imbalance ratio and the number of attributes still have a considerable influence on the Class Balance Accuracy. To achieve better results, it is necessary to determine a good sample. Determination of samples that are in the safe region gives quite good results. This is indicated by the Class Balance Accuracy value given by the Hybrid Approach with Distance Feature, which gives better results when compared to the results given by MultiRandBal.

The test results for Hamming Loss show the number of attributes that influence the results obtained most. Then the second most influential parameter after the number of attributes is the Balance Ratio. The greater the number of attributes and the greater the imbalance ratio, the greater the value of Hamming Loss, which means that the greater the misclassification that occurs. The lower the Hamming Loss value, the better the results obtained. The results showed that the difference between the Hybrid Approach with Distance Feature and MultiRandBal was not too big, but in general, the Hybrid Approach with Distance Feature method gave better results than MultiRandBal.

D. Statistical Tests

Based on the results of the tests carried out, it can be seen that generally, the results given by the Hybrid Approach with Distance Feature give better results when compared to MultiRandBal. It is interesting to test whether the differences are significant enough or not. For the purposes of the significance test, the Wilcoxon Signed-Rank Test is used [43]. Table V provides the results of the statistical test.

TABLE V
STATISTICAL TESTS USING WILCOXON SIGNED-RANK TEST

Performance Measurement	P-Value	Hypothesis
Augmented R-Value	0.0312500	Because the P-Value value is < 0.05, this indicates that the idea that the Augmented R-Value value obtained shows a significant difference is accepted. So it can be said that the overlapping handling results obtained by the Hybrid Approach with Distance Feature are better than the overlapping results obtained by MultiRandBal.
Class Average Accuracy	0.0312500	Similar to the results obtained by Augmented R-Value, the results obtained by Class Average Accuracy also show a significant difference that can be accepted (because the P-Value value is <0.05). This result generally clarifies that a good overlapping treatment can give positive results to the accuracy of the classification results.
Class Balance Accuracy	0.0625000	It is accepted that there are no appreciable differences between MultiRandBal and the Hybrid Approach with Distance Feature. (because P-Value>0.05)
Hamming Loss	0.0312500	The results of hypothesis testing indicate that when compared to MultiRandBal, the Hybrid Approach with Distance Feature exhibits a significant difference (because P-Value <0.05). These results indicate that the Hybrid Approach with Distance Feature can minimize classification errors that occur and this indicates that the results of handling multi-class imbalances provided by the Hybrid Approach with Distance Feature are generally better.

E. Discussion

This study shows 2 (two) different approaches to determining the oversampling process in SMOTE. The MultiRandBal approach uses random determination while the Hybrid Approach with Distance Feature is based on efforts to determine the best sample based on distance, with the main objective being samples originating in the Safe Region. It is expected that samples from the Safe Region can minimize the occurrence of overlapping, where high overlapping indicates a decrease in classification accuracy.

Testing on the results of handling multi-class imbalance in this study is based on the parameters of Augmented R-Value for Multi-Class, Class Average Accuracy, Class Balance Accuracy, and Hamming Loss. For these four parameters, the results given by the Hybrid Approach with Distance Feature are better than MultiRandBal. The test results of the

significance level of differences indicate significant differences in the parameters of Augmented R-Value for Multi-Class, Class Average Accuracy, and Hamming Loss.

In general, the imbalance ratio and the number of attributes greatly affect the Augmented R-Value for Multi-Class, Class Average Accuracy, and Class Balance Accuracy parameters. As for Hamming Loss, the number of attributes that affect the most is then followed by the imbalance ratio. This study's results indicate a direct relationship between the effect of overlapping and multi-class imbalance on the accuracy of the classification results. Overlapping is more often ignored when compared to class imbalance. However, it can be seen from the results of the study that the higher the overlap (which means that the overlap is more serious), the lower the accuracy of the classification results obtained.

The interesting thing is that for Class Balance Accuracy, although the results show that the Hybrid Approach with Distance Feature gives better results than MultiRandBal, the differences are insignificant. This can be understood because the Class Balance Accuracy is a balance between the accuracy of each existing class. The determination of the sample from the safe region tends to accommodate the handling of multi-class imbalance in the minority class.

IV. Conclusion

Based on the findings in Tables III, IV, and V, it is possible to conclude that both approaches have produced positive outcomes for handling multi-class imbalances. However, the results obtained by the Hybrid Approach with Distance Feature on several parameters are better. There are significant differences in the parameters of Augmented R-Value for Multi-Class, Class Average Accuracy, and Hamming Loss. As for the Class Balance Accuracy parameter, the difference obtained is not significant.

Implementing the Distance Feature in the Hybrid Approach for determining samples in safe regions has proven effective. In addition to dealing with multi-class imbalance problems, it can also handle overlapping. Thus, this study also shows a new approach to the oversampling process in SMOTE. It is hoped that this research can develop methods that can provide better accuracy results on datasets with a large number of attributes.

Acknowledgment

The Ministry of Education, Culture, Research, and Technology of Indonesia supported this study.

References

[1] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, and F. Herrera, "Dynamic ensemble selection for multi-class imbalanced datasets," *Information Sciences*, vol. 445–446, pp. 22–37, Jun. 2018, doi: 10.1016/j.ins.2018.03.002.

[2] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Machine Learning with Applications*, vol. 9, p. 100375, Sep. 2022, doi: 10.1016/j.mlwa.2022.100375.

[3] Y. Zhang, T. Sun, and C. Jiang, "Biomacromolecules as carriers in drug delivery and tissue engineering," *Acta Pharmaceutica Sinica B*, vol. 8, no. 1, pp. 34–50, Jan. 2018, doi: 10.1016/j.apsb.2017.11.005.

[4] X. Chao, G. Kou, Y. Peng, and A. Fernández, "An efficiency curve for evaluating imbalanced classifiers considering intrinsic data characteristics: Experimental analysis," *Information Sciences*, vol. 608, pp. 1131–1156, Aug. 2022, doi: 10.1016/j.ins.2022.06.045.

[5] P. Sadhukhan and S. Palit, "Adaptive learning of minority class prior to minority oversampling," *Pattern Recognition Letters*, vol. 136, pp. 16–24, Aug. 2020, doi: 10.1016/j.patrec.2020.05.020.

[6] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from Class-Imbalanced Data: Review of Methods and Applications," *Expert Systems With Applications*, vol. 73, pp. 220–239, May 2017.

[7] A. Zhang, H. Yu, Z. Huan, X. Yang, S. Zheng, and S. Gao, "SMOTE-RkNN: A hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors," *Information Sciences*, vol. 595, pp. 70–88, May 2022, doi: 10.1016/j.ins.2022.02.038.

[8] M. Koziarski, "Potential Anchoring for imbalanced data classification," *Pattern Recognition*, vol. 120, p. 108114, Dec. 2021, doi: 10.1016/j.patcog.2021.108114.

[9] Z. Chen, J. Duan, L. Kang, and G. Qiu, "A hybrid data-level ensemble to enable learning from highly imbalanced dataset," *Information Sciences*, vol. 554, pp. 157–176, Apr. 2021, doi: 10.1016/j.ins.2020.12.023.

[10] A. S. Desuky and S. Hussain, "An Improved Hybrid Approach for Handling Class Imbalance Problem," *Arab J Sci Eng*, vol. 46, no. 4, pp. 3853–3864, Apr. 2021, doi: 10.1007/s13369-021-05347-7.

[11] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Information Sciences*, vol. 512, pp. 1214–1233, Feb. 2020, doi: 10.1016/j.ins.2019.10.048.

[12] Q. Li, Y. Song, J. Zhang, and V. S. Sheng, "Multi-class imbalanced learning with one-versus-one decomposition and spectral clustering," *Expert Systems with Applications*, vol. 147, p. 113152, Jun. 2020, doi: 10.1016/j.eswa.2019.113152.

[13] T. R. Hoens, Q. Qian, N. V. Chawla, and Z.-H. Zhou, "Building Decision Trees for the Multi-class Imbalance Problem," in *Advances in Knowledge Discovery and Data Mining*, 2012, pp. 122–134.

[14] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognition*, vol. 57, pp. 164–178, Sep. 2016, doi: 10.1016/j.patcog.2016.03.012.

[15] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

[16] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *1*, vol. 61, pp. 863–905, Apr. 2018.

[17] J. Bi and C. Zhang, "An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme," *Knowledge-Based Systems*, vol. 158, pp. 81–93, Oct. 2018, doi: 10.1016/j.knosys.2018.05.037.

[18] M. S. Santos, P. H. Abreu, N. Japkowicz, A. Fernández, and J. Santos, "A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research," *Information Fusion*, vol. 89, pp. 228–253, Jan. 2023, doi: 10.1016/j.inffus.2022.08.017.

[19] H. K. Lee and S. B. Kim, "An overlap-sensitive margin classifier for imbalanced and overlapping data," *Expert Systems with Applications*, vol. 98, pp. 72–83, May 2018, doi: 10.1016/j.eswa.2018.01.008.

[20] X. Gao *et al.*, "A multi-class classification using one-versus-all approach with the differential partition sampling ensemble," *Engineering Applications of Artificial Intelligence*, vol. 97, p. 104034, Jan. 2021, doi: 10.1016/j.engappai.2020.104034.

[21] B. Chen, S. Xia, Z. Chen, B. Wang, and G. Wang, "RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise," *Information Sciences*, vol. 553, pp. 397–428, Apr. 2021, doi: 10.1016/j.ins.2020.10.013.

[22] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," *Intelligence-Based Medicine*, vol. 3–4, p. 100023, Dec. 2020, doi: 10.1016/j.ibmed.2020.100023.

[23] K. De Angeli *et al.*, "Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types," *Journal of Biomedical Informatics*, vol. 125, p. 103957, Jan. 2022, doi: 10.1016/j.jbi.2021.103957.

[24] E. R. Q. Fernandes and A. C. P. L. F. de Carvalho, "Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning," *Information Sciences*, vol. 494, pp. 141–154, Aug. 2019, doi: 10.1016/j.ins.2019.04.052.

[25] N. K. Mishra and P. K. Singh, "Feature construction and smote-based imbalance handling for multi-label learning," *Information Sciences*, vol. 563, pp. 342–357, Jul. 2021, doi: 10.1016/j.ins.2021.03.001.

[26] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Information Sciences*, vol. 542, pp. 92–111, Jan. 2021, doi: 10.1016/j.ins.2020.07.014.

[27] X. Tao *et al.*, "SVDD-based weighted oversampling technique for imbalanced and overlapped dataset learning," *Information Sciences*, vol. 588, pp. 13–51, Apr. 2022, doi: 10.1016/j.ins.2021.12.066.

[28] M. Koziarski, M. Woźniak, and B. Krawczyk, "Combined Cleaning and Re-sampling algorithm for multi-class imbalanced data with label noise," *Knowledge-Based Systems*, vol. 204, p. 106223, Sep. 2020, doi: 10.1016/j.knosys.2020.106223.

[29] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," *Artificial Intelligence in Medicine*, vol. 104, p. 101815, Apr. 2020, doi: 10.1016/j.artmed.2020.101815.

[30] Y. Liu, Y. Liu, B. X. B. Yu, S. Zhong, and Z. Hu, "Noise-robust oversampling for imbalanced data classification," *Pattern Recognition*, vol. 133, p. 109008, Jan. 2023, doi: 10.1016/j.patcog.2022.109008.

[31] J. J. Rodríguez, J.-F. Díez-Pastor, Á. Arnaiz-González, and L. I. Kuncheva, "Random Balance ensembles for multi-class imbalance learning," *Knowledge-Based Systems*, vol. 193, p. 105434, Apr. 2020, doi: 10.1016/j.knosys.2019.105434.

[32] P. Vuttipittayamongkol and E. Elyan, "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data," *Information Sciences*, vol. 509, pp. 47–70, Jan. 2020, doi: 10.1016/j.ins.2019.08.062.

[33] Q. Chen, Z.-L. Zhang, W.-P. Huang, J. Wu, and X.-G. Luo, "PF-SMOTE: A novel parameter-free SMOTE for imbalanced datasets," *Neurocomputing*, vol. 498, pp. 75–88, Aug. 2022, doi: 10.1016/j.neucom.2022.05.017.

[34] T. G.s., Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, and S. Chennupati, "An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets," *Machine Learning with Applications*, vol. 8, p. 100267, Jun. 2022, doi: 10.1016/j.mlwa.2022.100267.

[35] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, Jul. 2012, doi: 10.1109/TSMCC.2011.2161285.

[36] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, Jun. 2022, doi: 10.1016/j.jksuci.2022.06.005.

[37] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "A First Approach to Deal with Imbalance in Multi-label Datasets," in *Hybrid Artificial Intelligent Systems*, Berlin, Heidelberg, 2013, pp. 150–160. doi: 10.1007/978-3-642-40846-5_16.

[38] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behavioural Processes*, vol. 148, pp. 56–62, Mar. 2018, doi: 10.1016/j.beproc.2018.01.004.

[39] P. Branco, L. Torgo, and R. P. Ribeiro, "Relevance-Based Evaluation Metrics for Multi-class Imbalanced Domains," in *Advances in Knowledge Discovery and Data Mining*, Cham, 2017, pp. 698–710. doi: 10.1007/978-3-319-57454-7_54.

[40] L. Mosley, "A balanced approach to the multi-class imbalance problem," *Graduate Theses and Dissertations*, Jan. 2013, doi: https://doi.org/10.31274/etd-180810-3375.

[41] N. K. Mishra and P. K. Singh, "FS-MLC: Feature selection for multi-label classification using clustering in feature space," *Information Processing & Management*, vol. 57, no. 4, p. 102240, Jul. 2020, doi: 10.1016/j.ipm.2020.102240.

[42] A. Frank and A. Asuncion, "UCI Machine Learning Repository." University of California, School of Information and Computer Science, 2010. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.

[43] F. Wilcoxon, "Individual Comparisons by Ranking Methods on JSTOR," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.